level_one

```
%pyspark
from pyspark import SparkFiles
url = "https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Video_Games_v1_00.tsv.gz"
spark.sparkContext.addFile(url)
video_games_df = spark.read.csv(SparkFiles.get("amazon_reviews_us_Video_Games_v1_00.tsv.gz"), sep="\t", header=True, inferSchem
video_games_df.show()
```

SPARK JOBS FINISHED

```
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+---------
----+-----------+----+-----------------+--------------------+--------------------+-------------------+
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|product_category|star_rating|helpful_v
otes|total_votes|vine|verified_purchase|     review_headline|         review_body|        review_date|
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+---------
----+-----------+----+-----------------+--------------------+--------------------+-------------------+
|         US|   12039526| RTIS3L2M1F5SM|B001CXYMFS|     737716809|Thrustmaster T-Fl...|     Video Games|          5|
   0|          0|   N|                Y|an amazing joysti...|Used this for Eli...|2015-08-31 00:00:00|
|         US|    9636577| R1ZV7R40OLHKD|B00M920ND6|     569686175|Tonsee 6 buttons ...|     Video Games|          5|
   0|          0|   N|                Y|Definitely a sile...|Loved it,  I didn...|2015-08-31 00:00:00|
|         US|    2331478|R3BH071QLH8QMC|B0029CSOD2|      98937668|Hidden Mysteries:...|     Video Games|          1|
   0|          1|   N|                Y|            One Star|poor quality work...|2015-08-31 00:00:00|
|         US|   52495923|R127K9NTSXA2YH|B00GOOSV98|      23143350|GelTabz Performan...|     Video Games|          3|
   0|          0|   N|                Y|good, but could b...|nice, but tend to...|2015-08-31 00:00:00|
|         US|   14533949|R32ZWUXDJPW27Q|B00Y074JOM|     821342511|Zero Suit Samus a...|     Video Games|          4|
   0|          0|   N|                Y|   Great but flawed.|Great amiibo, gre...|2015-08-31 00:00:00|
|         US|    2377552|R3AQQ4YUKJWBA6|B002UBI6W6|     328764615|Psyclone Recharge...|     Video Games|          1|
   0|          0|   N|                Y|            One Star|The remote consta...|2015-08-31 00:00:00|
```

Took 8 min 45 sec. Last updated by anonymous at February 04 2019, 2:09:15 PM.

```
%pyspark
# Row Count
video_games_df.count()
```

SPARK JOB (http://172.17.0.2:4040/jobs/job?id=45) FINISHED

```
1785997
```

Took 31 sec. Last updated by anonymous at February 04 2019, 2:10:16 PM.

```
%pyspark
from pyspark.sql.functions import to_date
# Review DataFrame
review_id_df = video_games_df.select(["review_id", "customer_id", "product_id", "product_parent", to_date("review_date", 'yyyy-
```

SPARK JOB (http://172.17.0.2:4040/jobs/job?id=46) FINISHED

```
review_id_df.show()
```

level_one

```
+--------------+-----------+----------+--------------+-----------+
|     review_id|customer_id|product_id|product_parent|review_date|
+--------------+-----------+----------+--------------+-----------+
| RTIS3L2M1F5SM|   12039526|B001CXYMFS|     737716809| 2015-08-31|
| R1ZV7R40OLHKD|    9636577|B00M920ND6|     569686175| 2015-08-31|
|R3BH071QLH8QMC|    2331478|B0029CSOD2|      98937668| 2015-08-31|
|R127K9NTSXA2YH|   52495923|B00GOOSV98|      23143350| 2015-08-31|
|R32ZWUXDJPW27Q|   14533949|B00Y074JOM|     821342511| 2015-08-31|
|R3AQQ4YUKJWBA6|    2377552|B002UBI6W6|     328764615| 2015-08-31|
|R2F0POU5K6F73F|   17521011|B008XHCLFO|      24234603| 2015-08-31|
|R3VNR804HYSMR6|   19676307|B00BRA9R6A|     682267517| 2015-08-31|
| R3GZTM72WA2QH|     224068|B009EPWJLA|     435241890| 2015-08-31|
| RNQOY62705W1K|   48467989|B0000AV7GB|     256572651| 2015-08-31|
|R1VTIA3JTYBY02|     106569|B00008KTNN|     384411423| 2015-08-31|
|R29DOU8791QZL8|   48269642|B000A3IA0Y|     472622859| 2015-08-31|
|R15DUT1VIJ9RJZ|   52738710|B0053BQN34|     577628462| 2015-08-31|
|R3IMF2MQ3OU9ZM|   10556786|B002I0HIMI|     988218515| 2015-08-31|
|R23H79DHOZTYAU|    2963837|B0081FH12M|     770100932| 2015-08-31|
```

Took 1 sec. Last updated by anonymous at February 04 2019, 2:10:19 PM.

```
%pyspark
products_df = video_games_df.select(["product_id", "product_title"]).drop_duplicates()
```

FINISHED

Took 0 sec. Last updated by anonymous at February 04 2019, 2:10:21 PM.

```
%pyspark
reviews_df = video_games_df.select(["review_id", "review_headline", "review_body"])
reviews_df.show(10)
```

SPARK JOB (http://172.17.0.2:4040/jobs/job?id=47) FINISHED

```
+--------------+--------------------+--------------------+
|     review_id|     review_headline|         review_body|
+--------------+--------------------+--------------------+
| RTIS3L2M1F5SM|an amazing joysti...|Used this for Eli...|
| R1ZV7R40OLHKD|Definitely a sile...|Loved it,  I didn...|
|R3BH071QLH8QMC|            One Star|poor quality work...|
|R127K9NTSXA2YH|good, but could b...|nice, but tend to...|
|R32ZWUXDJPW27Q|   Great but flawed.|Great amiibo, gre...|
|R3AQQ4YUKJWBA6|            One Star|The remote consta...|
|R2F0POU5K6F73F|              A Must|I have a 2012-201...|
|R3VNR804HYSMR6|         Five Stars|Perfect, kids lov...|
| R3GZTM72WA2QH|         Five Stars|            Excelent|
| RNQOY62705W1K|         Four Stars|Slippery but expe...|
+--------------+-------------------+--------------------+
only showing top 10 rows
```

Took 1 sec. Last updated by anonymous at February 04 2019, 2:10:24 PM.

SPARK JOB (http://172.17.0.2:4040/jobs/job?id=48) FINISHED

```
%pyspark
customers_df = video_games_df.groupby("customer_id").agg({"customer_id": "count"}).withColumnRenamed("count(customer_id)", "cus
customers_df.show()
```

```
+-----------+--------------+
|customer_id|customer_count|
+-----------+--------------+
|   48670265|             1|
|   49103216|             2|
|    1131200|             1|
|   43076447|             2|
|   46261368|             1|
|    4883305|             5|
|   41192649|             1|
|   40985731|             7|
|   10437900|             2|
|   22245671|             1|
|    2574873|             1|
|    4696154|             1|
|    5621202|             1|
|    5871933|             2|
|   44089812|             1|
```

Took 33 sec. Last updated by anonymous at February 04 2019, 2:10:59 PM.

SPARK JOB (http://172.17.0.2:4040/jobs/job?id=49) FINISHED

```
%pyspark
vine_df = video_games_df.select(["review_id", "star_rating", "helpful_votes", "total_votes", "vine"])
vine_df.show(10)
```

```
+--------------+-----------+-------------+-----------+----+
|     review_id|star_rating|helpful_votes|total_votes|vine|
+--------------+-----------+-------------+-----------+----+
| RTIS3L2M1F5SM|          5|            0|          0|   N|
| R1ZV7R40OLHKD|          5|           0|          0|   N|
|R3BH071QLH8QMC|          1|           0|          1|   N|
|R127K9NTSXA2YH|          3|           0|          0|   N|
|R32ZWUXDJPW27Q|          4|           0|          0|   N|
|R3AQ04YUKJMBA6|          1|           0|          0|   N|
|R2F0POU5K6F73F|          5|           0|          0|   N|
|R3VNR804HYSMR6|          5|           0|          0|   N|
| R3GZTM72WA2QH|          5|           0|          0|   N|
| RNQOY62705W1K|          4|           0|          0|   N|
+--------------+-----------+------------+----------+----+
only showing top 10 rows
```

Took 1 sec. Last updated by anonymous at February 04 2019, 2:11:03 PM.

```
%pyspark
mode = "append"
jdbc_url="jdbc:postgresql://mydbinstance.cnhgk1ilyahu.us-east-2.rds.amazonaws.com:5432/my_data_class_db"
config = {"user":"root", "password": "CodingRocks!", "driver":"org.postgresql.Driver"}
```

FINISHED

Took 0 sec. Last updated by anonymous at February 04 2019, 2:24:13 PM.

```
%pyspark
# Write review_id_df to table in RDS
review_id_df.write.jdbc(url=jdbc_url, table='review_id_table', mode=mode, properties=config)
```

SPARK JOB (http://172.17.0.2:4040/jobs/job?id=60) FINISHED

Took 10 min 23 sec. Last updated by anonymous at February 04 2019, 2:35:38 PM.

```
%pyspark
# Write products_df to table in RDS
products_df.write.jdbc(url=jdbc_url, table='products', mode=mode, properties=config)
```

SPARK JOB (http://172.17.0.2:4040/jobs/job?id=61) FINISHED

Took 2 min 11 sec. Last updated by anonymous at February 04 2019, 2:37:57 PM.

```
%pyspark
```

```
# Write customers_df to table in RDS
customers_df.write.jdbc(url=jdbc_url, table='customers', mode=mode, properties=config)
```

SPARK JOB (http://172.17.0.2:4040/jobs/job?id=62) FINISHED

Took 5 min 40 sec. Last updated by anonymous at February 04 2019, 2:43:55 PM.

level_one

```
%pyspark
# Write vine_df to table in RDS
vine_df.write.jdbc(url=jdbc_url, table='vine_table', mode=mode, properties=config)
```

SPARK JOB (http://172.17.0.2:4040/jobs/job?id=63) FINISHED

Took 7 min 1 sec. Last updated by anonymous at February 04 2019, 2:52:25 PM.

```
%pyspark
```

READY